Spatial Approach to Pose Variations in Face Verification

Licesio J. Rodríguez-Aragón, Ángel Serrano, Cristina Conde, and Enrique Cabello

Universidad Rey Juan Carlos, c\ Tulipán, s/n,
E-28933, Móstoles, Madrid, Spain
{licesio.rodriguez.aragon, angel.serrano, cristina.conde, enrique.cabello}@urjc.es
http://frav.escet.urjc.es

Abstract. Spatial dimension reduction methods called Two Dimensional PCA and Two Dimensional LDA have recently been presented. These variations of traditional PCA and LDA consider images as 2D matrices instead of 1D vectors. The robustness to pose variations of these advances at verification tasks, using SVM as classification algorithm, is here shown.

The new methods endowed with a classification strategy of SVMs, seriously improve, specially for pose variations, the results achieved by the traditional classification of PCA and SVM.

1 Introduction

Some of the fields where biometrics play a relevant role are not only the improvement of security but also the development of smart environments where individuals are able to interact with computers in a human related way [1]. Dimensionality reduction is an important and necessary preprocessing of multidimensional data, as face images. Recent tests to measure the progress recently made towards face recognition show that accuracy on frontal face with indoor lighting goes beyond 90%, which is promising for early stages of recognition tasks [2]. On the other hand, face recognition among different pose or illumination is far from acceptable. Robustness to this changes in facial images is searched in many ways.

Analysis of the effects of pose [3] and illumination [4,5] variations over each face have been studied, searching for invariant characteristics or analyzing the perturbations introduced in the data. A normalization task is aimed by detecting characteristic points and measuring distances [6]. Three dimensional models of facial images are obtained through laser scanners [7], increasing the cost and the complexity of the problem. From our point of view, these methods improve the performance of the classification but traditional methods avoid dealing with an important problem, the spatial structure of the images.

Face recognition is different from classical pattern recognition, since there are many individual classes and only a few images per class. Dimension reduction

A.F. Famili et al. (Eds.): IDA 2005, LNCS 3646, pp. 351–361, 2005.

methods commonly used, like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA), and Gabor Filters [8], as well as other improved variations, like Independent Component Analysis (ICA) [9] and Kernel Principal Component Analysis (KPCA) [10], obtain a feature set for each image. Classical methods use vectorized representations of the images containing the faces instead of working with data in matrix representation. The main drawbacks of the classical vectorized projection methods is that it is easy to be subjected to gross variations and thus, high sensitive to any changes in pose, illumination etc.

New advances on feature extraction methods called Two-Dimensional Principal Component Analysis [11,12] and Two-Dimensional Linear Discriminant Analysis [13,14] have shortly been presented, and preliminar experiments and junctions of these new methods with SVM are the focus of this work. Experiments are performed over a wide set of subjects, joined in a facial database of images which allow the measurement of the advances of the recognition task to pose variations, specially to rotated faces.

2 Feature Extraction

Traditional feature extraction techniques require that 2D face images are vectorized into a 1D row vector to then perform the dimension reduction [8,9,10]. The resulting image vectors belong to a high-dimensional image vector space where covariance matrices are evaluated with a high associated computational cost.

Recently, a Two-Dimensional PCA method (2DPCA) and Two-Dimensional LDA (2DLDA) have been developed for bidimensional data feature extraction. Both methods are based on 2D matrices rather than 1D vectors, preserving spatial information.

2.1 Principal Component Analysis

Given a set of images $I_1, I_2, \ldots, I_N$ of height h and width w, PCA considers the images as 1D vectors in a $h \cdot w$ dimensional space. The facial images are projected onto the eigenspace spanned by the leading ortonormal eigenvectors, those of higher eigenvalue, from the sample covariance matrix of the training images. Once the set of vectors has been centered, the sample covariance matrix is calculated, resulting a matrix of dimension $h \cdot w \times h \cdot w$. It is widely known that if $N \ll h \cdot w$, there is no need to obtain the eigenvalue decomposition of this matrix, because only N eigenvectors will have a non zero associated eigenvalue [15]. The obtention of these eigenvectors only requires the decomposition of an $N \times N$ matrix, considering as variables the images, instead of the pixels, and therefore considering pixels as individuals.

Once the first d eigenvectors are selected and the proportion of the retained variance fixed (Fig. 1), $\sum_1^d \lambda_i / \sum_1^N \lambda_i$, being $\lambda_1 > \lambda_2 > \cdots > \lambda_N$ the eigenvalues, a projection matrix A is formed with $h \cdot w$ rows and d columns, one for each eigenvector. Then a feature vector $Y_{d\times 1}$ is obtained as a projection of each image $I_{h \cdot w \times 1}$, considered as a 1D vector, onto the new eigenspace.

2.2 Linear Discriminant Analysis

The previous method maximizes the total scatter retained by the fixed dimension. Information provided by the labels of the set of images, $I_1, I_2, \ldots, I_N$, is not used. Linear Discriminant Analysis shapes the scatter in order to make it more reliable for classification. Traditional Linear Discriminant Analysis uses this information to maximize between-class scatter whereas within-class scatter is minimized simplifying the classification process and focusing the problem in a more reliable way.

As images are transformed into a 1D vector, the method faces the difficulty that the within-class scatter matrix, of dimension $h \cdot w \times h \cdot w$, is always singular as the number of images N of the set is usually much lower than the number of pixels in an image. An initial projection using PCA is done to a lower dimensional space so that the within-scatter matrix is non singular. Then applying the standard Fisher Linear Discriminant Analysis, the dimension is finally reduced [16].

2.3 Two-Dimensional Principal Component Analysis

The consideration of images $I_{h \times w}$ as 1D vectors instead as 2D structures is not the right approach to retain spatial information. Pixels are correlated to their neighbours and the transformation of images into vectors produces a loss of information preserving the dimensionality. On the contrary, the main objective of these methods is the reduction of dimensionality and the least loss of information as possible.

The idea recently presented as a variation of traditional PCA, is to project an image $I_{h \times w}$ onto X^{PCA} by the following transformation [11,12],

$$Y_{h \times 1} = I_{h \times w} \cdot X^{PCA}_{w \times 1}. \tag{1}$$

As result, a h dimensional projected vector Y, known as projected feature vector of image I, is obtained. The total covariance matrix S_X over the set of projected feature vectors of training images $I_1, I_2, \ldots, I_N$ is considered. The mean of all the projected vectors, $\overline{Y} = \overline{I} \cdot X^{PCA}$, being $\overline{I}$ the mean image of the training set, is taken into account.

$$\begin{aligned} S_X &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (Y_i - \overline{Y})(Y_i - \overline{Y})^T \\ &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} [(I_i - \overline{I})X][(I_i - \overline{I})X]^T \end{aligned} \tag{2}$$

The maximization of the total scatter of projections is chosen as the criterion to select the vector X^{PCA}. The total scatter of the projected samples is characterized by the trace of the covariance matrix of the projected feature vectors. Applying the criterion to (2) the following expression is obtained,

$$J(X) = tr(S_X) = X^T [\frac{1}{N} \sum_{i=1}^{N} (I_i - \overline{I})^T (I_i - \overline{I})] X. \tag{3}$$

What is known as image covariance matrix S defined as a $w \times w$ nonnegative matrix can be directly evaluated using the training samples,

$$S = \frac{1}{N} \sum_{i=1}^{N} (I_i - \overline{I})^T (I_i - \overline{I})]. \tag{4}$$

The optimal projection axis X^{PCA} is the unitary vector that maximizes (3), which corresponds to the eigenvector of S of largest associated eigenvalue.

2.4 Two-Dimensional Linear Discriminant Analysis

The idea presented as 2DPCA, has been upgraded to consider the class information [13,14]. Suppose there are L known pattern clases having M samples for each class, $N = L \cdot M$. The idea is to project each image as in (1), but to obtain X^{LDA} with the information provided by the classes. The covariance over the set of images can be decomposed into between-class and within-class. The mean of projected vectors as in 2DPCA as well as the mean of projected vectors of the same class $\overline{Y^j} = \overline{I^j} \cdot X^{LDA}$, being $\overline{I^j}$ the mean image of the class $j = 1, \ldots, L$, are taken into account.

$$\begin{aligned} S_{XB} &= \textstyle\sum_{j=1}^{L} M(\overline{Y^j} - \overline{Y})(\overline{Y^j} - \overline{Y})^T \\ &= \textstyle\sum_{j=1}^{L} M[(\overline{I^j} - \overline{I})X][(\overline{I^j} - \overline{I})X]^T \end{aligned} \tag{5}$$

$$\begin{aligned} S_{XW} &= \textstyle\sum_{j=1}^{L} \sum_{i=1}^{M} (Y_i^j - \overline{Y^j})(Y_i^j - \overline{Y^j})^T \\ &= \textstyle\sum_{j=1}^{L} \sum_{i=1}^{M} [(I_i^j - \overline{I^j})X][(I_i^j - \overline{I^j})X]^T \end{aligned} \tag{6}$$

The objective function maximized in this case to select X^{LDA} is considered a class specific linear projection criterion, and can be expressed as

$$J(X) = \frac{tr(S_{XB})}{tr(S_{XW})}. \tag{7}$$

The total between and within covariances are defined as $w \times w$ nonnegative matrices and can be directly evaluated.

$$S_B = \sum_{j=1}^{L} M[(\overline{I^j} - \overline{I})][(\overline{I^j} - \overline{I})]^T; \quad S_W = \sum_{j=1}^{L} \sum_{i=1}^{M} [(I_i^j - \overline{I^j})][(I_i^j - \overline{I^j})]^T \tag{8}$$

Both matrices are formally identical to the corresponding traditional LDA, and by maximizing (7) the within-class scatter is minimized whereas the between-class scatter is maximized, giving as result the maximization of discriminating information. The optimal projection axis X^{LDA} is the unitary vector that maximizes (7), which corresponds to the eigenvector of $S_B \cdot S_W^{-1}$, of largest associated eigenvalue.

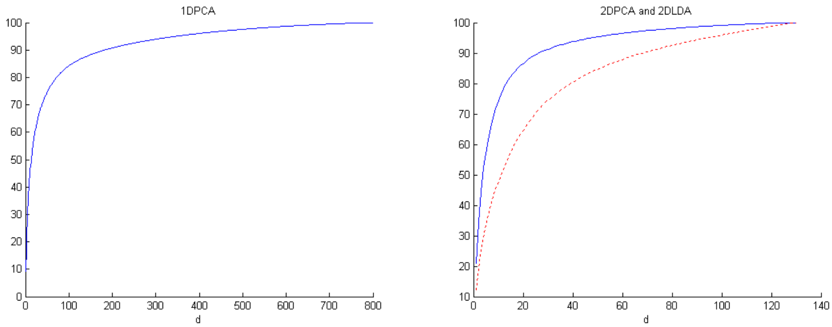

Fig. 1. Evolution of the retained variance percentage for the dimension reduction methods. Left, PCA, for $N = 800$ possible dimensions. Right, 2DPCA in solid line and 2DLDA in dashed line, for $w = 130$ possible dimensions.

3 Projection and Reconstruction

As in traditional PCA, a proportion of retained variance is fixed in 2DPCA and 2DLDA (Fig. 1), $\sum_1^d \lambda_i / \sum_1^w \lambda_i$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_w$ are the eigenvalues and $X_1, X_2, \ldots, X_d$ are the eigenvectors corresponding to the d largest eigenvalues.

Once d is fixed, $X_1, X_2, \ldots, X_d$ are the ortonormal axes used to perform the feature extraction. Let $V = [Y_1, Y_2, \ldots, Y_d]$ and $U = [X_1, X_2, \ldots, X_d]$, then

$$V_{h \times d} = I_{h \times w} \cdot U_{w \times d}. \tag{9}$$

A set of projected vectors, $Y_1, Y_2, \ldots, Y_d$, are obtained for both methods. Each projection over an optimal projection vector is a vector, instead of a scalar as in traditional PCA. A feature matrix $V_{h \times d}$ for each considered dimension reduction method is produced, containing either the most amount of variance, or the most discriminating features of image I.

3.1 Image Reconstruction

In this dimension reduction methods, a reconstruction of the images from the features is possible. An approximation of the original image with the retained information determined by d is obtained.

$$\begin{aligned} \widetilde{I}_{h \cdot w \times 1} &= A_{h \cdot w \times d} \cdot Y_{d \times 1} \quad \text{PCA image reconstruction.} \\ \widetilde{I}_{h \times w} &= V_{h \times d} \cdot U^T_{d \times w} \quad \text{2DPCA or 2DLDA image reconstruction.} \end{aligned} \tag{10}$$

4 Classification with SVM

SVM is a method of learning and separating binary classes [17], it is superior in classification performance and is a widely used technique in pattern recognition and especially in face verification tasks [18].

Given a set of features $y_1, y_2, \ldots, y_N$ where $y_i \in \mathbb{R}^n$, and each feature vector associated to a corresponding label $l_1, l_2, \ldots, l_N$ where $l_i \in \{-1, +1\}$, the aim of a SVM is to separate the class label of each feature vector by forming a hyperplane

$$(\omega \cdot y) + b = 0, \quad \omega \in \mathbb{R}^n, b \in \mathbb{R}. \tag{11}$$

The optimal separating hyperplane is determined by giving the largest margin of separation between different classes. This hyperplane is obtained through a minimization process subjected to certain constrains. Theoretical work has solved the existing difficulties of using SVM in practical application [19].

As SVM is a binary classifier, a *one vs. all* scheme is used. For each class, each subject, a binary classifier is generated with positive label associated to feature vectors that correspond to the class, and negative label associated to all the other classes.

4.1 Facial Verification Using SVM

In our experiments a group of images from every subject is selected as the training set and a disjoint group of images is selected as the test set. The training set is used in the feature extraction process through PCA, 2DPCA and 2DLDA. Then, the training images are projected onto the new ortonormal axes and the feature vector (PCA), or vectors (2DPCA,2DLDA), are obtained. For each subject the required SVMs are trained.

Several strategies have been used to train and combine the SVMs. When training and classifying PCA features, each image generates one feature vector $Y_{d\times 1}$ and one SVM is trained for each subject, with its feature vectors labelled as $+1$ and all the other feature vectors as -1.

On the other hand, for feature vectors obtained from 2DPCA and 2DLDA, each image generates a set of projected vectors, $V_{h\times d} = [Y_1, Y_2, \ldots, Y_d]$, and three different strategies have been considered. First strategy generates a unique feature vector through a concatenation of the d projected vectors, then one SMV is trained for each subject as in PCA. The second and third approaches consider the d projected vectors and consequently for each subject d SVMs are trained, one for each feature vector. These d outputs are then combined to produce a final classification output, first through an arithmetic mean and secondly trough a weighted mean.

Once the SVMs are trained, images from the test set are projected onto the eigenspace obtained from the training set. The features of the test set are classified through the SVMs to measure the performance of the generated system.

For the SVM obtained from the PCA and from the concatenation strategy of 2DPCA and 2DLDA feature vectors, the output is compared with the known label of every test image. However, for the ensemble of SVMs obtained from the 2DPCA and 2DLDA feature vectors, the d outputs are combined whether through an arithmetic or a weighted mean. Arithmetic approach combines the d outputs through an arithmetic mean. At weighted approach, every output is weighted with the amount of variance explained by its dimension, that means that each output will be taken in account proportionally to the value of the eigenvalue associated to the corresponding eigenvector: $\lambda_i / \sum_{j=1}^{d} \lambda_j$ is the weight for the $i-$SVM, $i = 1, 2, \ldots, d$.

To measure the system performance a cross validation procedure is carried out. Results are then described by using Receiver Operating Curve, ROC curve, as there are four possible experiment outcomes: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The system threshold can then be adjusted to more or less sensitiveness, but in order to achieve fewer errors new and better methods, like 2DPCA and 2DLDA, are required.

5 Design of Experiment

The Face Recognition and Artificial Vision[1] group (FRAV) at the Universidad Rey Juan Carlos, has collected a quite complete set of facial images for 109

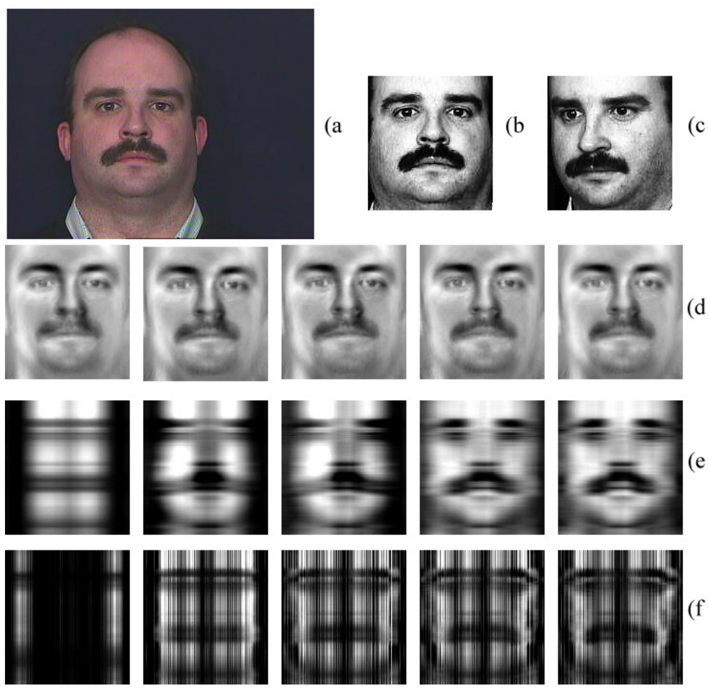

Fig. 2. a) One of the original frontal images in the FRAV2D database. b) Automatically selected window containing the facial expression of the subject in equalized gray scale. c) Sample of a pose variation face, rotated 15°, used to evaluate the performance of the verification. d) From left to right, reconstructed images (10), for $d = 10, 50, 90, 150, 170$, from PCA projection. e) and f) From left to right, reconstructed images (10), for $d = 1, 2, 3, 4, 5$, from 2DPCA and 2DLDA projections respectively.

[1] http://frav.escet.urjc.es

subjects. All the images have been taken under controlled conditions of pose and illumination. A partial group of this database is freely available for research purposes.

The images are colored and of size 240×320 pixels with homogeneous background color. A window of size 140×130 pixels containing the most meaningful part of the face, has been automatically selected in every image and stored in equalized gray scale. That is the information that will be analyzed through the dimension reduction and classification methods (Fig. 2).

The purpose of the following experiments is to confront the robustness to pose variations of the traditional PCA method and classifying strategies to the new proposed 2DPCA and 2DLDA methods in the task of face verification through SVM. Each experiment has been performed for 100 randomly chosen subjects from the whole FRAV2D. In all the experiments, the train set for the extraction of the feature vectors and for the classifiers training is formed by eight frontal images of each subject. Then, the classifiers have been tested over four $15°$ rotated images to measure the performance of the system at pose variations.

Different tests for the reduced dimension of the projections with different values have been carried out. Results for the best performance of each method are presented as ROC curves (Fig. 3), showing the compared performance of the verification process using PCA, 2DPCA and 2DLDA. True positive rate (TP), that is the proportion of correct classifications to positive verification problems, and true negative rate (TN), that is the proportion of correct classifications to negative verification problems, are plotted. Besides, the equal error rate (EER), that is the value for which false positive rate (FP) is equal to false negative rate (FN), is presented for each experiment that has been undertaken (Fig. 4).

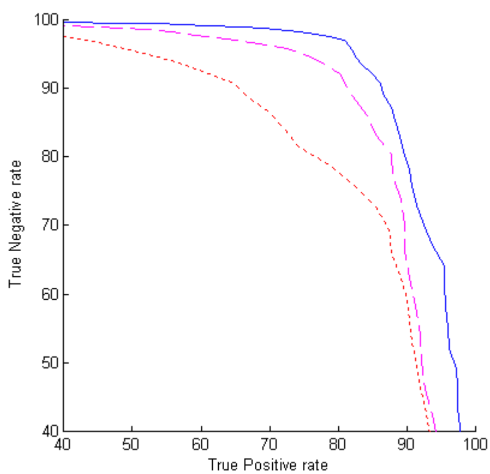

Fig. 3. ROC curves for the best performance of each dimension reduction method, with TP rate in abscises and TN rate in ordinates. The performance of PCA with $d = 170$ in dotted line, 2DLDA with $d = 2$ under concatenated strategy in dashed line and 2DPCA with $d = 1$ in solid line.

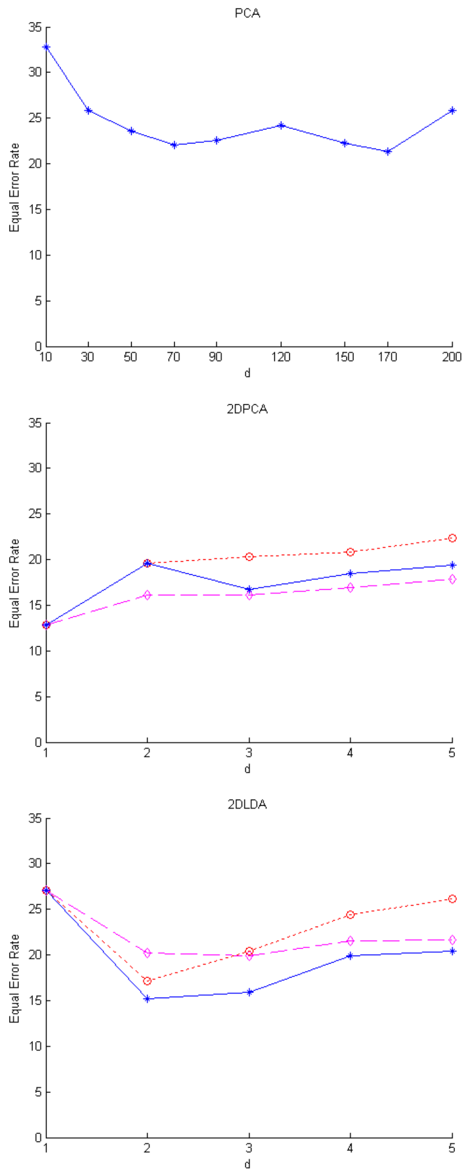

Fig. 4. Top, Equal Error Rate for PCA dimension reduction method for different values of d. Best performance is done for $d = 170$, $EER = 21.33\%$. Center, Equal Error Rate for 2DPCA dimension reduction method for different values of d and the three SVM strategies, concatenated in solid line, arithmetic mean in dotted line and weighted mean in dashed line. Best performance is done for $d = 1$, $EER = 12.89\%$. Bottom, Equal Error Rate for 2DLDA dimension reduction method for different values of d, as in the previous figure the three strategies have been considered. Best performance is done for $d = 2$, $EER = 15.19\%$ with concatenated strategy.

6 Conclusions

Best results are achieved for the spatial reduction method 2DPCA, as presented at the the ROC curves and the EER values for each method (Fig. 3, 4). Improvements are over 8% with respect to PCA.

Both spatial methods improve the performance of traditional PCA but serious differences appear. 2DPCA reaches its maximum accuracy at $d = 1$, while 2DLDA needs $d = 2$ to reach its best performance, both being quite low from $w = 130$ possible dimensions. PCA reaches its best performance at $d = 170$ from $N = 800$ possible dimension. None of the three classifying strategies are able to improve the results while increasing the dimension at 2DPCA. 2DLDA best performance is reached with concatenation strategy, though weighted mean strategy, as in 2DPCA, seems more robust to the increase of dimension. Spatial methods lead to an eigenvector decomposition of matrices with sizes, $w \times w$, much smaller than PCA, $N \times N$.

It is clear that the spatial dimension reduction methods are more reliable for the purpose of face verification, specially for pose variations (Fig. 2), but deeper work has to be done to use all the information provided by the dimension reduction methods in order to achieve a more accurate verification.

Acknowledgments

Authors would like to thank César Morales García for his enthusiastic work. Also thanks must be given to every one that offered his help to join FRAV2D data base. This work has been partially supported by URJC grant GVC−2004 − 04.

References

1. Bowyer, K. W.: Face recognition technology: security versus privacy. IEEE Technology and society magazine. **Spring** (2004) 9–20
2. Messer, K., Kittler, J., Sadeghi, M., et al. : Face authentification test on the BANCA database. Proceedings of the International Conference on Pattern Recognition. (2004) 523–532.
3. Gross, R., Jie Yang and Waibel, A.: Growing Gaussian mixture models for pose invariant face recognition. In Proceedings. 15th International Conference on Pattern Recognition. (2000) 1088–1091.
4. Batur, A.U. and Hayes, M.H.I.I.I.: Linear subspaces for illumination robust face recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2001) vol. 2 296–301.
5. Chen, H., Belhumeur, P. and Jacobs, D.: In search of Illumination Invariants. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2000) 254–261.
6. Conde, C., Cipolla, R., Rodríguez-Aragón, L.J., Serrano Á. and Cabello, E.: 3d facial feature location with spin images. In Proceedings of the 9th International Association for Pattern Recognition Conference on Machine Vision Applications. (2005) 418–421.

7. Lu, X., Colbry, D. and Jain, A. K.: Three-Dimensional Model Based Face Recognition. In Proceedings of International Conference on Pattern Recognition. (2004) 362–366.
8. Pang, S., Kim, D. and Bang, S. Y.: Membership authentication in the dynamic group by face classification using SVM ensemble. Pattern Recognition Letters **24** (2003) 215–225.
9. Kim, T., Kim, H., Hwang, W. and Kittler, J.: Independent Component Analysis in a local facial residue space for face recognition. Pattern Recognition. **37** (2004) 1873–1885.
10. Cao L.J. and Chong W.K.: Feature extraction in support vector machine: a comparsion of PCA, KPCA and ICA. Proceedings of the International Conference on Neural Information Processing. Vol. 2 (2002) 1001–1005.
11. Yang, J. and Yang, J.: From image vector to matrix: a straightforward image projection technique–IMPCA vs. PCA. Pattern Recognition **35** (2002) 1997–1999.
12. Yang, J., Zhang, D., Frangi and F., Yang, J.: Two-Dimmensional PCA: A new approach to apperance-based face representation and recognition. IEEE Transacctions on Pattern Recognition and Machine Intelligence. **26** (2004) 131–137.
13. Li, M., Yuan, B.Z.: A novel statistical linear discriminant analysis for image matrix: two-dimensional fisherfaces. Proceedings of the International Conference on Signal Processing. (2004) 1419–1422.
14. Chen S., Zhu Y., Zhang D. and Yang J.: Feature extraction approaches based on matrix pattern: MatPCA and MatFLDA. Pattern Recognition Letters. In press.
15. Turk, M. and Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neurosicience. **3** (1991) 71–86.
16. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 711–720.
17. Cortes, C. and Vapnik, V.: Support vector network. Machine Learning. **20** (1995) 273–297.
18. Fortuna, J. and Capson, D.: Improved support vector classification using PCA and ICA feature space modiffication. Pattern Recognition **37** (2004) 1117–1129
19. Joachims, T.: Making large scale support vector machine learning practical. In: Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge, MA.